PRML Assignment 3

Rongali Rohith (EE19B114) Santosh G (EE19B055)

I. Aim

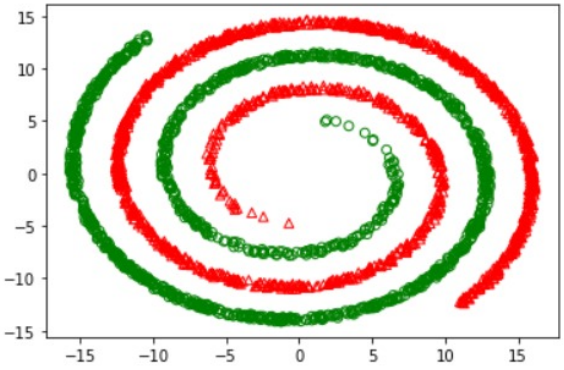
- To build a model to classify Image Data and Synthetic Data by considering class conditional data as Gaussian Mixture Model

II. Introduction

- Synthetic Data: The data belongs to two categories, which can separated by a non-linear decision boundary. The boundary is being computed after modelling the data and their probabilities into Gaussian Mixture Models.
- The data is separated into various clusters of roughly the same size and each data cluster is modelled as a normal distribution.

III. Code

- The data is separated into K-Clusters, and mean is computed for each cluster. Initial values are assigned to various Parameters such as Covariance Matrix, mixture popularity etc.
- The following is the initial data given, the data has been plotted to understand the complexity of the data and to estimate the number of clusters to be made



- The values are parsed through the Gaussian function to update the parameters. The process is repeated for several iterations such that the parameters converge to a final value.
- Gaussian Distributions are constructed, based on the finally computed means and variances and are finally mixed to fit the data in.
  The classes have been divided into two different plots. The data has been divided into K (= 20) clusters, mean of each cluster has been computed and has been plotted.

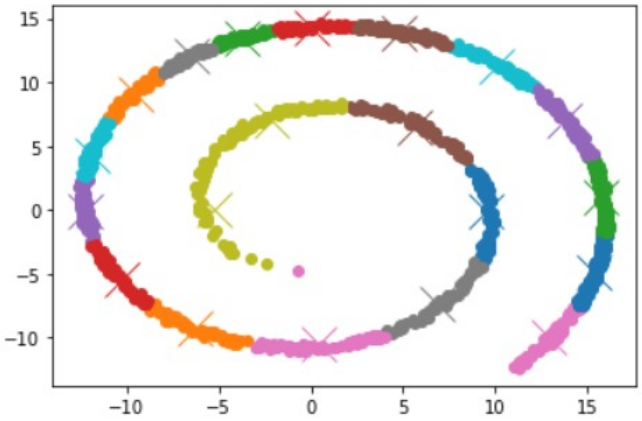


Fig. 1. Class 1 and corresponding K (=20) clusters and means

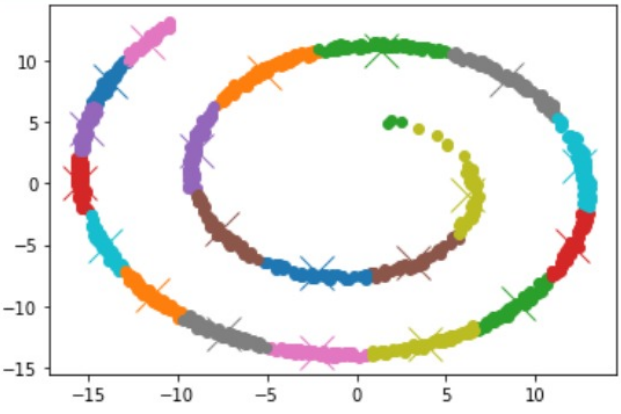


Fig. 2. Class 2 and corresponding K (=20) clusters and means

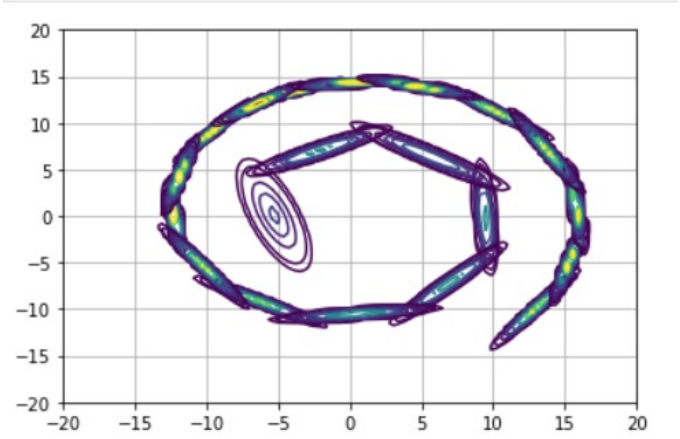


Fig. 3. Class 1 and corresponding K (=20) Gaussians

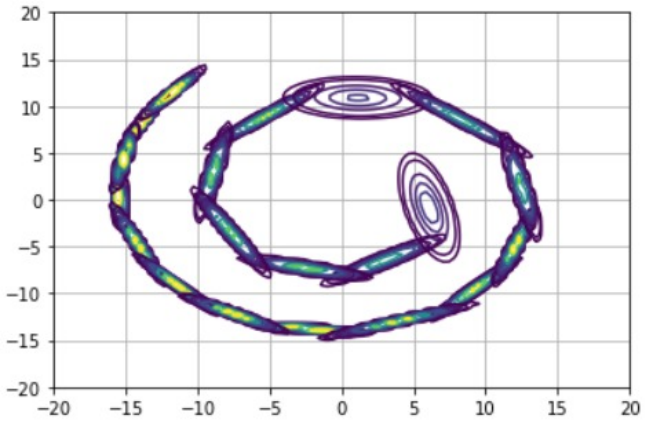


Fig. 4. Class 2 and corresponding K (=20) Gaussians

- To classify the development data, the maximum of probabilities of those points belonging to each of the 2K Gaussians (K from each class) has been considered and the data is hence classified accordingly.
- The following are the **Confusion matrices**, which can be used to analyze the mis-classifications and the accuracy based on the values.

Confusion matrix for GMM (Bayes, K = 20):

	Belongs to Class 1	Belongs to Class 2
Classified as Class 1	500	0
Classified as Class 2	1	499

The diagonal elements in the above confusion matrix is the count of data points that have been rightly classified, the non-diagonal elements represent the mis-classification

We can observe an accuracy of 99.9% i.e only 1 in 1000 has been mis-classified.

We can observe some aberrations in few Gaussians (i.e. being spread out), this can be seen due to the nature of the data in the cluster. The data either has outliers (can be seen in the case of top Gaussian in Class 2) or the data belongs to a curved cluster (can be seen in the case of central gaussians in both the classes) We have also used Naive Bayes (Diagonal Covariance matrix) to make cluesters and to classify the data. The following are the results obtained:

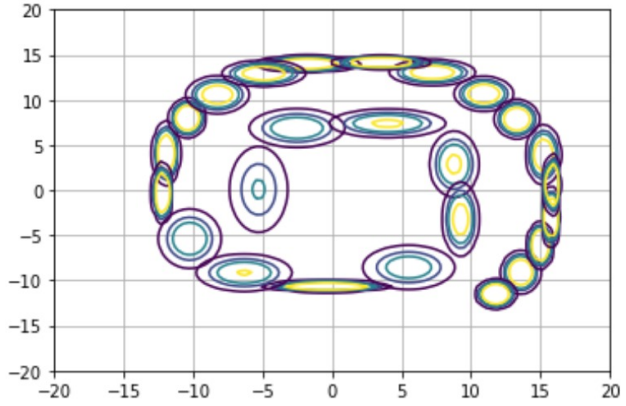


Fig. 5. Class 1 and corresponding K (=20) Gaussians

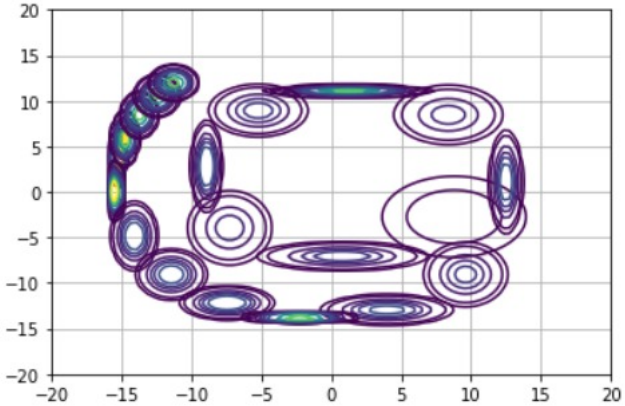


Fig. 6. Class 2 and corresponding K (=20) Gaussians

Confusion matrix for GMM (Naive Bayes, K = 20):

	Belongs to Class 1	Belongs to Class 2
Classified as Class 1	492	3
Classified as Class 2	8	497

The error rate is around 1.1% and is hence less accurate compared to the previous one.

Now, we have computed the Decision boundary and have included below: The Purple region corresponds to Class 1

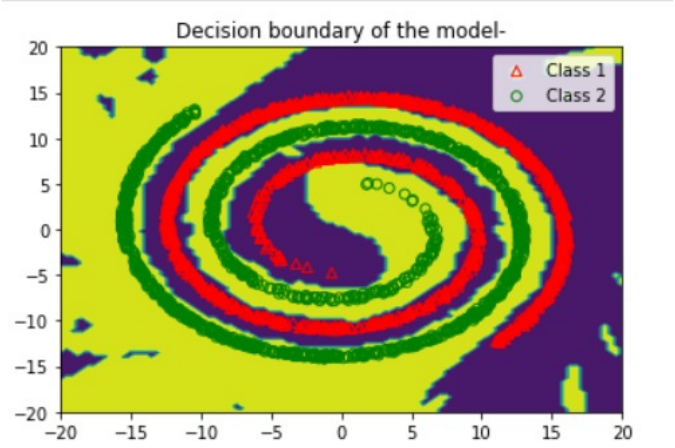


Fig. 7. Decision Boundary with data points for K=75

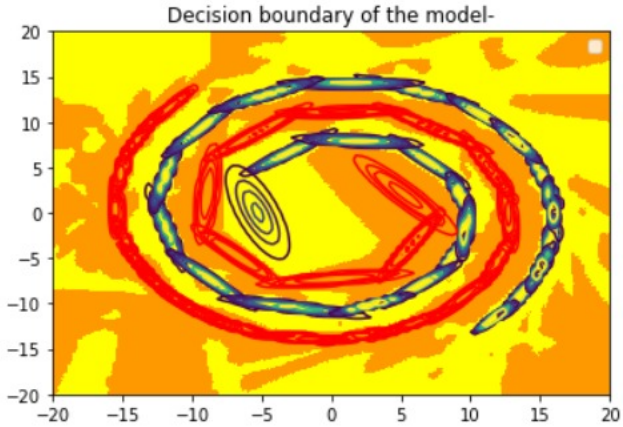


Fig. 8. Decision Boundary with Gaussians for K=20

where as the yellow region corresponds to Class 2 i.e the data in those regions are most likely to be classified/belong to the corresponding class.

Combing the decision boundary with the gaussians, we get the following plot:

We can also compare the performance of the model for various values of K, from the following ROC and DET curves.

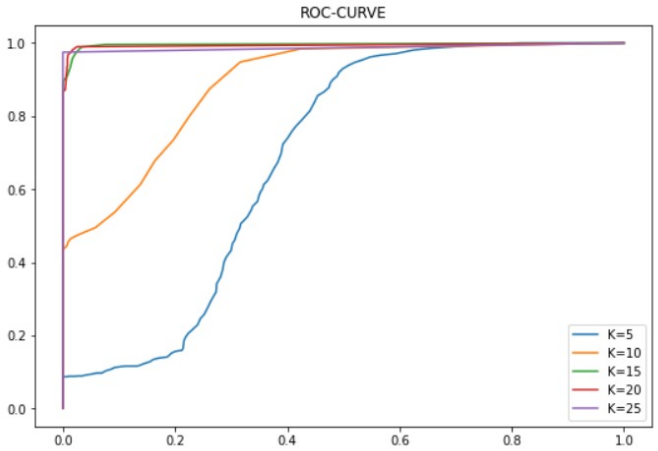


Fig. 9. ROC for various values of K

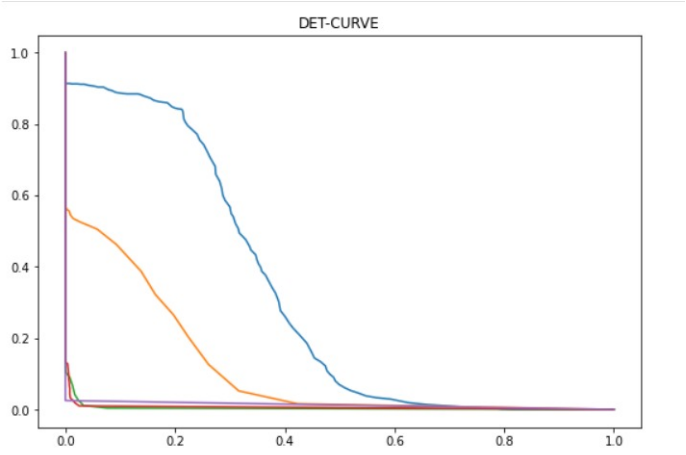


Fig. 10. DET for various values of K

From the above ROC plot it is clear that, the accuracy increases with the increase in value of K, and the accuracy is almost 100% for K=25.

IV. Image Data set

Image data set contains images belonging to 5 different classes. Each Image is broken into 36 blocks. Every image has 23 dimensional feature vectors. Processing the whole image at once would give a covariance matrix of very high dimension, 828x828. To avoid this, the image is divided into 6x6 blocks and each block is analysed

We trained the GMM model for each class using the function we've written before. While testing on the development data we took the log-likelihood i.e. sum of log(probability) summed over all the 36 blocks of the development data sample and assigned the score for each class.

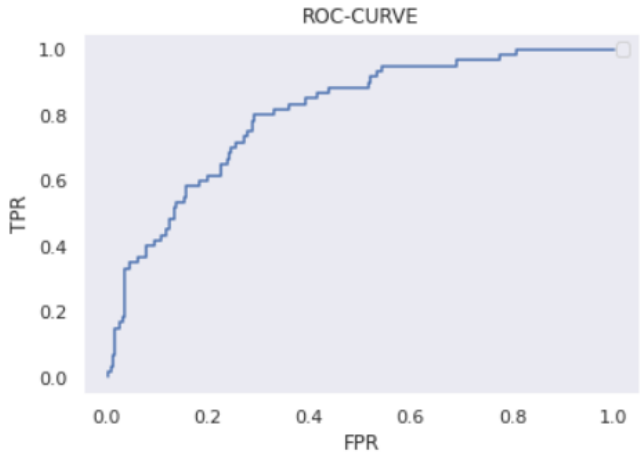


Fig. 11. ROC Curve for Image Dataset (K=30)

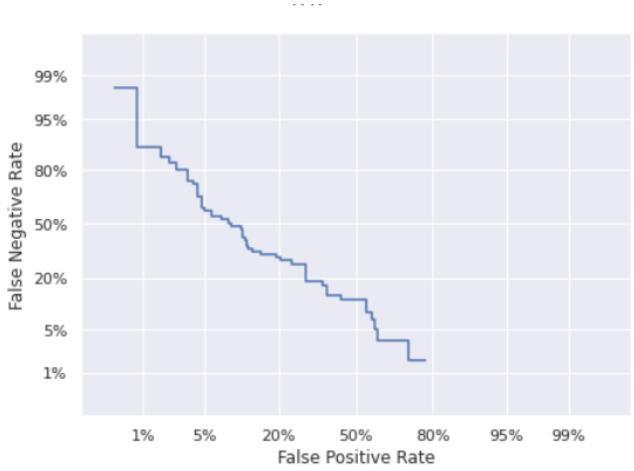


Fig. 12. DET Curve for Image Dataset (K=30)

We have experimented the same with various values of K and obtained the following results, as depicted in the ROC Plots below. We have achieved a best accuracy of 63% for value of K = 30.

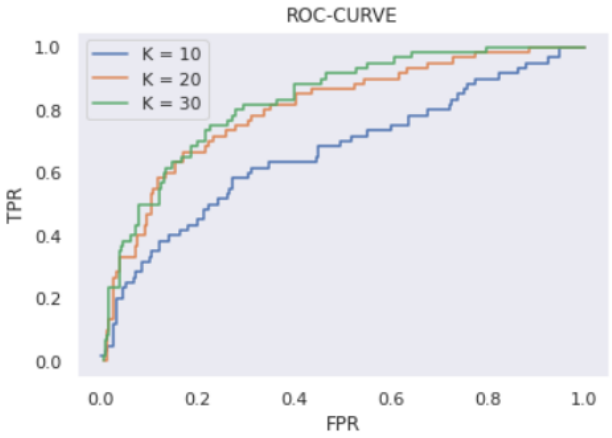


Fig. 13. ROC Curves for variosu values of K

For higher values of K, the computation was taking very long time and hence, many experiments involving higher values of K haven't been done.

PRML Assignment 3

Rongali Rohith (EE19B114) Santosh G (EE19B055)

I. **Aim**

- To build HMM Model and perform DTW on
  - a) Isolated spoken digit data set
  - b) Online Handwritten-Character data set
- To plot the ROC, DET plots and confusion matrices to draw inferences about model accuracy, working etc

II. **Data**

Speech Recordings of various numbers have been provided, from which the features also have been extracted. Similarly for Handwritten data, coordinates of the characters have been provided.

Specific to our team we have been given the following data:

Speech Recording Feature of numbers: **1, 3, 5, 9, z (Zero)**

Handwritten Character Data set of letters: **a, ai, bA, cha, lA**

The following is an image of the data provided.

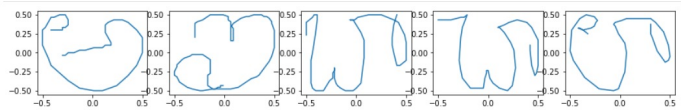


Fig. 1. Various Letters given to classify

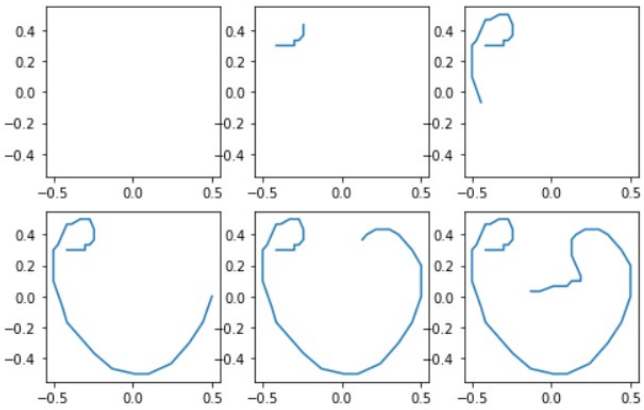


Fig. 2. Letter

DTWs;

- Dynamic Time Warping is performed on development data and the currently available training data.Based on the results, the development data is classified.

  Each data point in development data is tested against all the training-data to find the distance.Top K-Smallest distances are averaged, to find the score for each class.
  The classification is done based on these scores.

  The Hand Written data has been re-centered and normalised, to make it independent of the position, range and to increase the consistency among the data. As it can be seen in the above picture, the range has been shifted to [-0.5 to 0.5].

- After the classification of data, ROC curve has been plotted.
- We achieved an accuracy of 98% accuracy by performing DTW on Handwritten Character Data.(k=10)
- The following are the **Confusion matrices**, which can be used to analyze the mis-classifications and the accuracy based on the values.

Confusion matrix for DTW on Handwritten Character Data:

	Belongs to Class 1	Belongs to Class 2	Belongs to Class 3	Belongs to Class 4	Belongs to Class 5
Classified as Class 1	20	0	0	0	0
Classified as Class 2	0	20	0	0	0
Classified as Class 3	0	0	19	1	0
Classified as Class 4	0	0	1	19	0
Classified as Class 5	0	0	0	0	20

It can be seen that the error is occurring in Class 3 and Class 4, i.e. of 'bA' and 'chA'. These letters have high similarity, and hence the model is making mis-classifications.

The diagonal elements in the above confusion matrix is the count of data points that have been rightly classified, the non-diagonal elements represent the mis-classification

ROC and DET Curves:

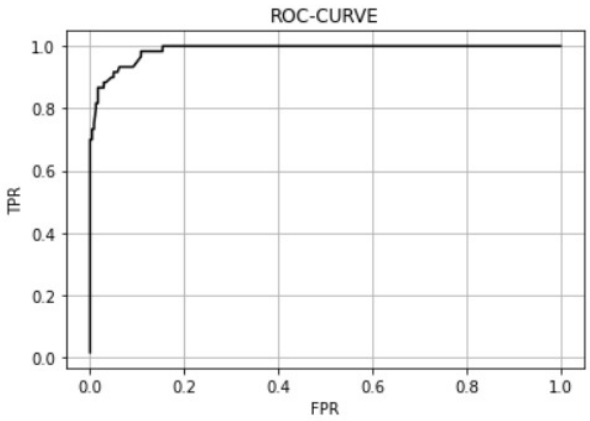


Fig. 3. ROC Curve for DTW of Online Handwritten-Character data set

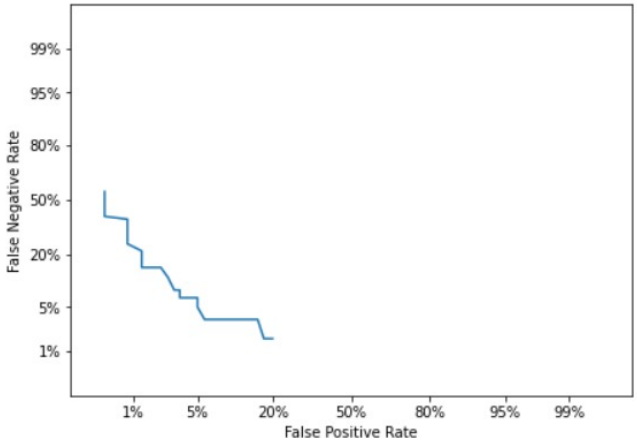


Fig. 4. DET Curve for DTW of Online Handwritten-Character data set

The data of spoken digits has been processed similarly and we have obtained the following results:

We achieved an accuracy of 96.667% while performing DTW on spoken digits.(k=15)
Confusion matrix for DTW on Isolated Spoken Digit Data:

	Belongs to Class 1	Belongs to Class 2	Belongs to Class 3	Belongs to Class 4	Belongs to Class 5
Classified as Class 1	12	0	1	1	0
Classified as Class 2	0	12	0	0	0
Classified as Class 3	0	0	11	0	0
Classified as Class 4	0	0	0	11	0
Classified as Class 5	0	0	0	0	12

ROC Curves:

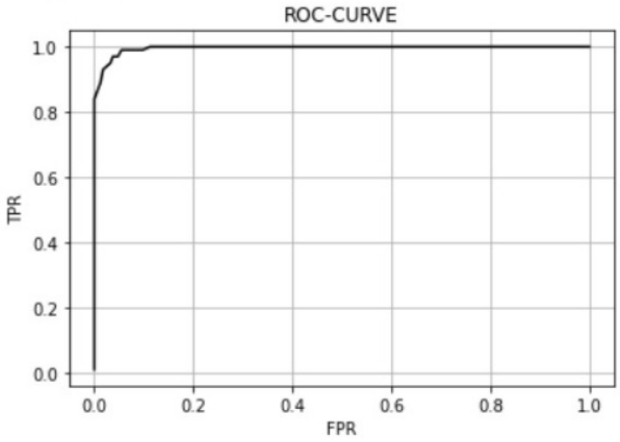


Fig. 5. ROC Curve for DTW of Isolated spoken digit data set

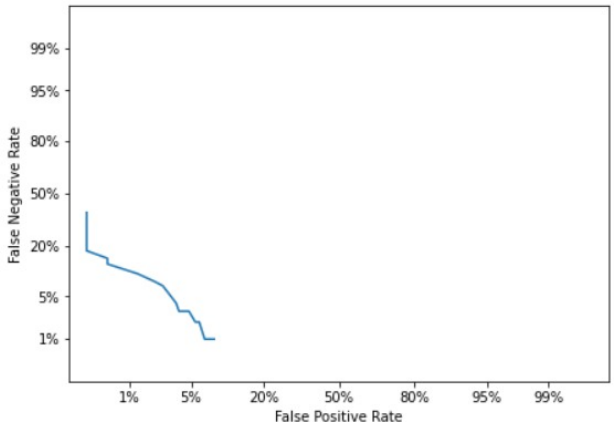


Fig. 6. DET Curve for DTW of Isolated spoken digit data set

HMMs:

For a given set of data we tried to tune the number of states and symbols to give the highest accuracy while classifying data.

We initially performed clustering to map the features to symbols.Using the training symbol sequences we trained a hmm model for each class.

Then on the development data we compute the likelihood of the observed sequences given each HMM model which is used for classification.

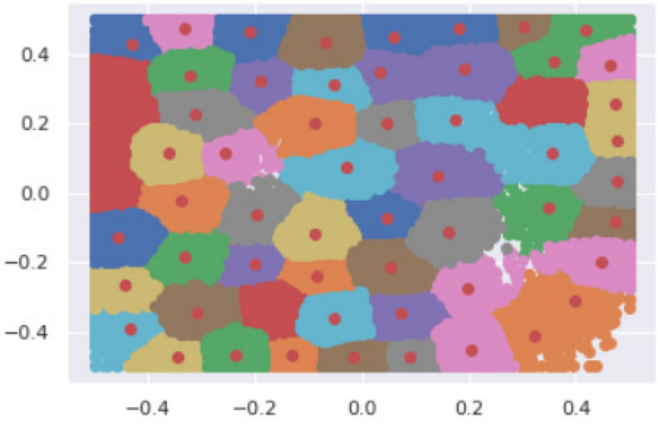


Fig. 7. Means after clustering for hand-written data

We got the highest accuracy for the following set of parameters(from whatever we could try):

1) Isolated digits: states = [5,5,5,5,6], no of symbols = 25
2) Online Hand-written: states = [10,10,11,11,10], no of symbols =60

ROC Curves:

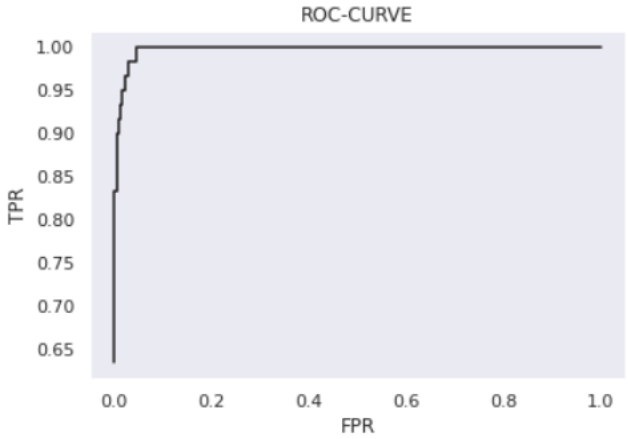


Fig. 8. ROC Curve for HMM of Isolated spoken digit data set

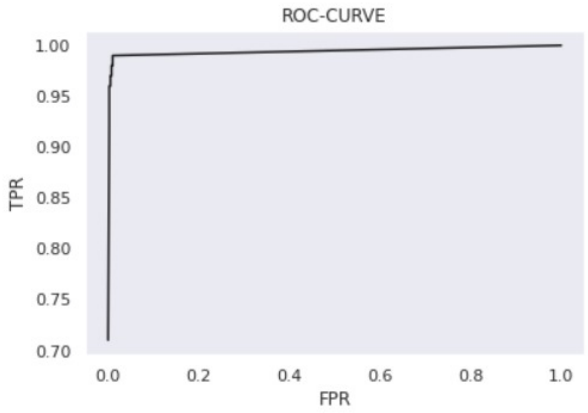


Fig. 9. ROC Curve for HMM of Online Handwritten-Character data set

Confusion matrix for HMM on Isolated Spoken Digit Data:

	Belongs to Class 1	Belongs to Class 2	Belongs to Class 3	Belongs to Class 4	Belongs to Class 5
Classified as Class 1	10	1	1	0	0
Classified as Class 2	0	12	0	0	0
Classified as Class 3	1	0	11	0	0
Classified as Class 4	1	0	0	11	0
Classified as Class 5	0	0	0	0	12

Confusion matrix for HMM on Handwritten Character Data:

	Belongs to Class 1	Belongs to Class 2	Belongs to Class 3	Belongs to Class 4	Belongs to Class 5
Classified as Class 1	20	0	0	0	0
Classified as Class 2	0	20	0	0	0
Classified as Class 3	0	0	19	1	0
Classified as Class 4	0	0	2	18	0
Classified as Class 5	0	0	0	0	20

Much like in the DTW case, here too misclassification occurs between bA and chA.

Also note that the number of symbols required for hand-written characters is higher as evident from the ROC plot in next page.

The following are the ROC and DET curves based on the results of classification for various values of K (Number of clusters).

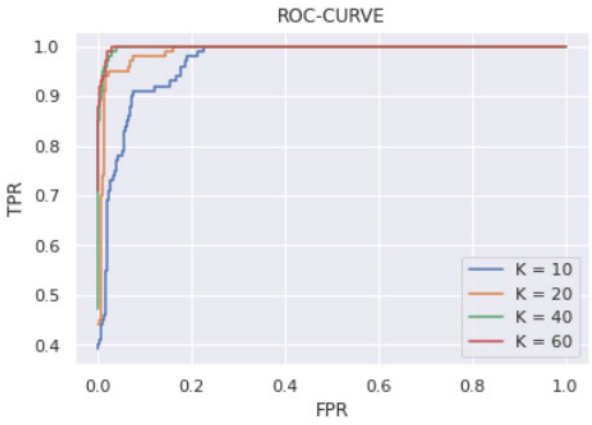


Fig. 10. ROC Curves for HMM of Handwritten Character Data

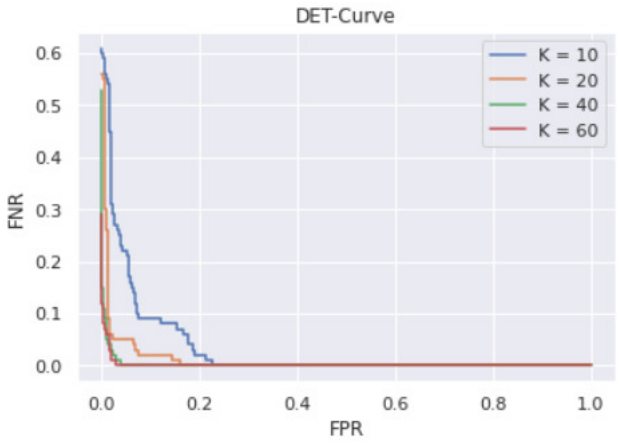


Fig. 11. DET Curves for HMM of Handwritten Character Data